An Advanced Data Fabric Architecture Leveraging Homomorphic Encryption and Federated Learning

by

Sakib Anwar Rieyan
19101024
Md. Raisul Kabir News
19101058
A.B.M. Muntasir Rahman
19101042
Sadia Afrin Khan
19101034
Sultan Tasneem Jawad Zaarif
19101206

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

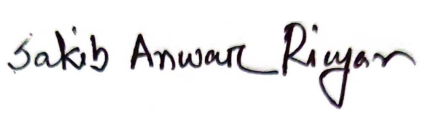
Department of Computer Science and Engineering
School of Data and Sciences
Brac University
March 2023

Declaration

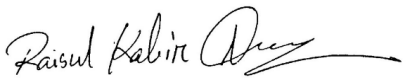
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material that has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

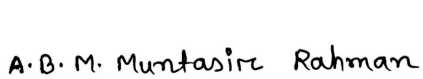
Student's Full Name & Signature:



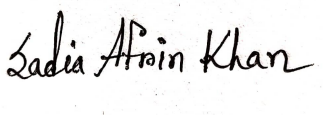
Sakib Anwar Rieyan
19101024



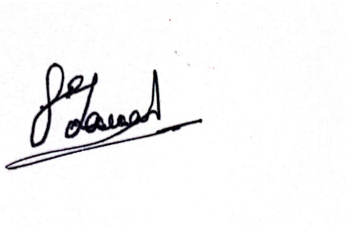
Md. Raisul Kabir News
19101058



A.B.M. Muntasir Rahman
19101042



Sadia Afrin Khan
19101034



Sultan Tasneem Jawad Zaarif
19101206

Approval

The thesis/project titled "An Advanced Data Fabric Architecture Leveraging Homomorphic Encryption and Federated Learning" submitted by

1. Sakib Anwar Rieyan (19101024)
2. Md. Raisul Kabir News (19101058)
3. A.B.M. Muntasir Rahman (19101042)
4. Sadia Afrin Khan (19101034)
5. Sultan Tasneem Jawad Zaarif (19101206)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on March 31, 2023.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, Ph.D.
Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, Ph.D.
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, Ph.D.
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

In this study, we present a novel approach for securely analyzing medical images using federated learning and partially homomorphic encryption within a distributed data fabric architecture. Our approach allows multiple parties to collaboratively train a machine learning model without exchanging raw data, while still maintaining compliance with laws and regulations such as HIPAA and GDPR. We demonstrate the effectiveness of our approach using pituitary tumor classification as a case study, achieving an overall accuracy of 83.31%. However, the primary focus of our work is on the development and evaluation of federated learning and partially homomorphic encryption as tools for secure medical image analysis. Our results show the potential for these techniques to be applied in other privacy-sensitive domains and contribute to the growing body of research on secure and privacy-preserving machine learning.

Keywords: Data Fabric, Federated Learning, Partially Homomorphic Encryption

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption.

Secondly, to our supervisor Md. Golam Rabiul Alam Sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally, our parents without their support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

Declaration i

Approval ii

Abstract iv

Acknowledgment v

Table of Contents vi

List of Figures viii

List of Tables ix

Nomenclature x

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Research Problems . . . 2
- 1.3 Contributions . . . 3

2 Background Studies 4
- 2.1 Data Fabric . . . 4
  - 2.1.1 Vanilla Architecture of Data Fabric . . . 5
- 2.2 Federated Learning . . . 8
- 2.3 Homomorphic Encryption . . . 9

3 Related Works 11

4 Methodology 18
- 4.1 Data Description . . . 18
- 4.2 Preprocessing Data . . . 19
- 4.3 Proposed Model . . . 19
  - 4.3.1 Advanced Architecture of Data Fabric . . . 19
  - 4.3.2 Implemented Federated Learning Framework . . . 23
  - 4.3.3 Comparative Analysis with Related Works . . . 24
  - 4.3.4 Model Specification . . . 25

5 Result Analysis 29

6 Conclusion **37**
6.1 Limitations 37
6.2 Future Works 38

Bibliography **42**

List of Figures

2.1 Vanilla Architecture of Data Fabric 7
2.2 Federated Learning Model 8
2.3 Homomorphic Encryption 10

4.1 Collected dataset from Kaggle [37] 18
4.2 Overview of the encrypted dataset 19
4.3 Accessing Data 20
4.4 Data Lifecycle Management 20
4.5 Exposing Data 21
4.6 Proposed Architecture of Data Fabric 22
4.7 Local Federated Learning Model 23
4.8 Global Federated Learning Model 24
4.9 VGG16 Model Architecture 25
4.10 VGG19 Model Architecture 26
4.11 ResNet50 Model Architecture 27
4.12 ResNet152 Model Architecture 27
4.13 Custom CNN Model Architecture 28

5.1 Performance of tested models on unencrypted and encrypted dataset 29
5.2 Confusion Matrix of VGG16 30
5.3 Training History of VGG16 31
5.4 Confusion Matrix of VGG19 31
5.5 Training History of VGG19 32
5.6 Confusion Matrix of ResNet50 32
5.7 Training History of ResNet50 33
5.8 Confusion Matrix of ResNet152 33
5.9 Training History of ResNet152 34
5.10 Confusion Matrix of CNN 34
5.11 Training History of Custom CNN 35
5.12 Receiver Operating Characteristic of Different Models 35
5.13 Receiver Operating Characteristic of Custom CNN 36

List of Tables

2.1 Comparison between the types of Homomorphic Encryption 9

4.1 Comparative Analysis 25

5.1 Accuracy scores across different models 30

Nomenclature

The next list describes several symbols & abbreviations that will be later used within the body of the document.

β Beta

AI Artificial intelligence

$APIs$ Application Programming Interfaces

CNN Convolutional Neural Network

ECC Elliptic Curve Cryptography

$EHRs$ Electronic Health Records

$GDPR$ General Data Protection Regulation Act

GNB Gaussian Naive Bayes

GPU Graphics Processing Unit

$HIPAA$ Health Insurance Portability and Accountability Act

IT Information Technology

MLE Maximum Likelihood Estimation

$MLOps$ Machine Learning Operations

MRI Magnetic Resonance Imaging

PCA Principal Component Analysis

PHE Partially Homomorphic Encryption

SVM Support Vector Machine

USL Ubiquitous Street Light

$VMware$ Virtual Machine ware

Chapter 1

Introduction

Artificial intelligence (AI) has become an integral part of our daily lives, including in the field of healthcare. In order to make the most of AI in healthcare, it is important to have access to large amounts of high-quality data. However, the confidentiality and sensitivity of healthcare data pose significant challenges to its storage and analysis. The volume of healthcare data is often very large, particularly due to the prevalence of image-based data. In addition, the confidentiality of healthcare data is of the utmost importance, as it is often personal and sensitive in nature.

To address these challenges, we have built an advanced data fabric architecture that brings together healthcare centers in a region and stores patient data and diagnoses in a secure and privacy-preserving manner. Our approach utilizes federated learning and partially homomorphic encryption to allow for collaborative machine learning on encrypted data, while still maintaining compliance with laws and regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [1] and the General Data Protection Regulation (GDPR) Act 2018 [2].

In this study, we have used pituitary tumor classification as a case study, employing various deep-learning models such as VGG16, VGG19, ResNet50, and ResNet152. Our results show promising potential for the use of federated learning and partially homomorphic encryption in secure medical image analysis. Specifically, we achieved good performance with VGG16 and VGG19 models, while ResNet50 and ResNet152 achieved lower accuracy and precision for both classes. However, our custom CNN architecture outperformed all of these pre-trained models in almost every metric that we used. Our findings contribute to the growing body of research on secure and privacy-preserving machine learning and demonstrate the potential for these techniques to be applied in other privacy-sensitive domains.

1.1 Motivation

In the field of medical image analysis, ensuring the security of sensitive patient data is of utmost importance. However, with the increasing use of machine learning and deep learning techniques for medical image analysis, there is a pressing need for an effective and secure architecture to handle such data. Previous studies have shown that the use of conventional security measures, such as encryption and access control, is not enough to ensure the privacy of patient data in the context of machine

learning and deep learning operations. In addition, the use of traditional centralized architectures for processing medical image data can be slow and resource-intensive, which can further compromise the security of the data.

Therefore, there is a clear need for a new, advanced data fabric architecture that is specifically designed to handle the unique challenges of securing medical image data while also supporting efficient machine learning and deep learning operations. This research aims to address this gap in the current state of the art by proposing and evaluating a novel architecture that is capable of effectively securing medical image data while also enabling fast and accurate machine learning and deep learning operations.

1.2 Research Problems

The integration of data into healthcare has the potential to improve the prediction of diseases and epidemics, enhance treatment outcomes, and prevent premature deaths. However, the confidentiality of healthcare data and the complexity of managing large and diverse datasets pose significant challenges to the integration of data into healthcare. Ensuring data security and privacy is of utmost importance, as security breaches in healthcare are on the rise. In a study on patient surveys and data security [3], it was found that 94% of hospitals had at least one security breach in the past two years. In addition, the high volume of healthcare data can make it difficult to effectively process, store, and communicate, and traditional methods may not be sufficient to handle the scale of the data. In one proposed solution [4], a big data healthcare cloud would host clinical, financial, social, physical, and psychological data from patients in a centralized location. However, proper governance of the data cloud is necessary to effectively work with and analyze complex data.

In this study, we aim to address these challenges by proposing an advanced data fabric architecture that brings together healthcare centers in a region and stores patient data and diagnoses in a secure and privacy-preserving manner using federated learning and partially homomorphic encryption. We demonstrate the effectiveness of our approach using pituitary tumor classification as a case study. However, the primary focus of our work is on the development and evaluation of federated learning and partially homomorphic encryption as tools for secure medical image analysis in the healthcare sector.

The research question that this study aims to answer is: How effective and practical is the implementation of advanced data fabric architecture using federated learning and partially homomorphic encryption for secure medical image analysis in the healthcare sector?

1.3 Contributions

Through our work, we show a fully-fledged data fabric architecture based on healthcare data can be built whilst complying with privacy regulations and maintaining good accuracy scores. Our main contribution is threefold:

- We propose an advanced data fabric architecture for storing and collaborating healthcare data in an encrypted form by using Partial Homomorphic Encryption (PHE) and sharing it with other parties without revealing its content. In this architecture, medical images of various patients/clients are encrypted on the client side and the encrypted images are trained and then collected. Further processing of data is done in its encrypted state. Here, The raw data was encrypted and generated to local weights using the local FL Model before getting global attention. Therefore even if the data was backtracked the end result will produce nothing but an encrypted image. Thus, this architecture provides a secure and efficient mechanism for processing encrypted data, while preserving data privacy and confidentiality regulations such as HIPAA and GDPR.
- The thesis proposes a federated learning framework that enables multiple clients to collaboratively train machine learning models on their respective data and then store the local updates, unlike the existing general federated learning frameworks which function on real-time local and global updates. This allows us access to modify, scale, merge or select local model updates before using them in the global model. In this way, the framework we proposed facilitates the systematic exchange of model updates between local and global models.
- Additionally, we customized a convolutional neural network (CNN) similar to that of VGG16 and VGG19 but having a lesser input size which contributes to a lesser parameter size compared to the mentioned models. This helps us to achieve more efficiency by reducing computational complexity as we are using Partially Homomorphic Encryption (PHE).

We further evaluate the proposed approach by implementing a prototype of the homomorphic encryption-based data fabric and the federated learning framework. The assessment indicates that the suggested method offers an effective and reliable means for sharing and analyzing data securely. The experimental results demonstrate that the proposed approach achieves satisfactory accuracy in the collaborative training of machine learning models, even when the data is encrypted.

Chapter 2

Background Studies

2.1 Data Fabric

According to Gartner [5], Data Fabric is a unified and integrated platform that enables data discovery, integration, management, and access across multiple environments. It provides a consistent and scalable approach to managing data assets that are distributed across various locations, such as on-premises, cloud, and edge computing. Data Fabric helps organizations to simplify and optimize their data management processes, reduce data silos, and enable real-time access to data. It also supports the creation of a self-service data marketplace, allowing users to discover, share, and consume data in a secure and governed manner. Data Fabric is increasingly becoming a critical component of modern data architectures, as organizations seek to manage the growing volume, velocity, and variety of data generated by digital business initiatives.

In our research, we are utilizing homomorphic encryption to classify pituitary tumors from MRI images in our dataset. We have used a Data Fabric architecture to store the weights of different machine-learning models as encrypted data. The ML models are run on client PCs, and the resulting encrypted data is saved in our Data Lake. Using homomorphic encryption, a server PC can perform computations on the encrypted data, allowing for the creation of a homogeneous global model. The server can then provide users with the requested results without compromising the privacy of the MRI images. This approach benefits from the Data Fabric's ability to provide a unified and integrated platform that enables data discovery, integration, management, and access across multiple environments. By utilizing homomorphic encryption and a Data Fabric architecture, we can classify pituitary tumors from MRI images in a privacy-preserving manner, contributing to the development of more secure and privacy-preserving medical imaging technologies.

2.1.1 Vanilla Architecture of Data Fabric

1. **Accessing Data:**

   (a) **Data Collecting and Encryption:** Data is a volatile resource, and Medical Data is considered highly sensitive as it can contain personally identifiable information such as names, addresses, dates of birth, and medical records which can be exploited if they fell into wrong hands. To comply with this shortcoming, in this architecture, data is neither collected nor stored in a central server which may possess the risk of data leakage.

   Here, firstly, to ensure privacy and reduce data volatility, medical data from various users are first selected and then encrypted with Partially Homomorphic Encryption (PHE). Subsequently, the encrypted data is selected to train the model locally for collecting updated model weights. The data of each user is generated and stored locally, without being transferred to the central server. Instead, the generated model updates are stored and merged for the global model formation.

   (b) **Master Data Management:** After the generation of local model updates and merging merged data is a feature selected for optimization and efficiency. Using feature selection, the important features or weights are selected which can help reduce the dimensionality of data, making it easier to work. It also reduces the risk of overfitting, improves model accuracy, and results in low computational cost and more efficiency due to lesser data to be trained.

   FedMax, FedAvg, and FedMin are optimization algorithms used in Federated Learning for feature selection. In all three algorithms, updated model weights are sent to the server/stored for future usage. However, in the case of FedMax, the server/user selects the model with the highest accuracy while it chooses the lowest loss model for FedMin and an average of all models for FedAvg.

   In our architecture, we selected FedMax as our feature selection algorithm to select important and relevant model weights as it showed more accuracy and efficiency compared to FedAvg and FedMin. The selected data is collected and kept together as "Master Data", with which we will continue our next works.

2. **Managing Life Cycle:**

   (a) **Governance:** Data governance is an essential component of data fabric architecture. Data fabric architecture is an approach to data management that enables organizations to manage and process data from multiple sources, locations, and formats. It provides a unified view of data

across the organization and supports various data processing requirements, such as data integration, analytics, and artificial intelligence.

Data governance in data fabric architecture refers to the policies, processes, and standards that organizations implement to manage their data assets effectively. Data governance helps organizations ensure that their data is accurate, consistent, and compliant with regulatory requirements. It also helps organizations manage data privacy, security, and access.

The following are some key considerations for data governance in data fabric architecture:

Data quality: Data governance policies should include measures to ensure data quality, such as data profiling, data cleansing, and data validation.

Metadata management: Data governance policies should include metadata management to ensure that data is properly tagged, categorized, and classified. Metadata helps organizations understand the meaning and context of their data and facilitates data discovery and reuse.

Data privacy and security: Data governance policies should include measures to ensure data privacy and security, such as access controls, data encryption, and data masking.

Data lineage: Data governance policies should include data lineage to track the origin, transformation, and movement of data across the organization. Data lineage helps organizations understand how data is used and facilitates compliance with regulatory requirements.

Data ownership and stewardship: Data governance policies should define data ownership and stewardship to ensure that data is managed and maintained by the appropriate individuals and teams.

(b) **Compliance:** Data compliance refers to adhering to relevant laws, regulations, and industry standards related to the handling, processing, and storage of data. In the context of data fabric, data compliance refers to ensuring that data is managed in accordance with these requirements across the entire data fabric. To ensure data compliance in a data fabric, it is necessary to establish policies and procedures that cover the entire data lifecycle, from data ingestion to archival and deletion. Personal data must be collected, processed, and stored in compliance with privacy regulations such as GDPR, CCPA, HIPAA, etc.

In this architecture, the feature selected weights were trained on various models such as VGG16, VGG19, ResNet 50, ResNet 152, etc. and updates are stored in a data lake structurally based on models they were trained on complying with HIPAA regulations.

3. **Exposing Data:** Data exposing refers to making data available for consumption and analysis by users or applications within an organization. Exposing data in a data fabric involves providing access to the data for authorized users or applications. There are several ways to expose data in a data fabric, including:

   APIs: Application Programming Interfaces (APIs) enable applications to access and retrieve data from the data fabric.

   Data Catalogs: A data catalog provides a searchable inventory of data assets in the data fabric. Users can discover and access data assets through the data catalog.

   Self-Service Analytics: A self-service analytics platform enables users to create their own queries and reports using the data available in the data fabric.

   Data Virtualization: Data virtualization enables users to access and combine data from multiple sources as if it were in a single location.

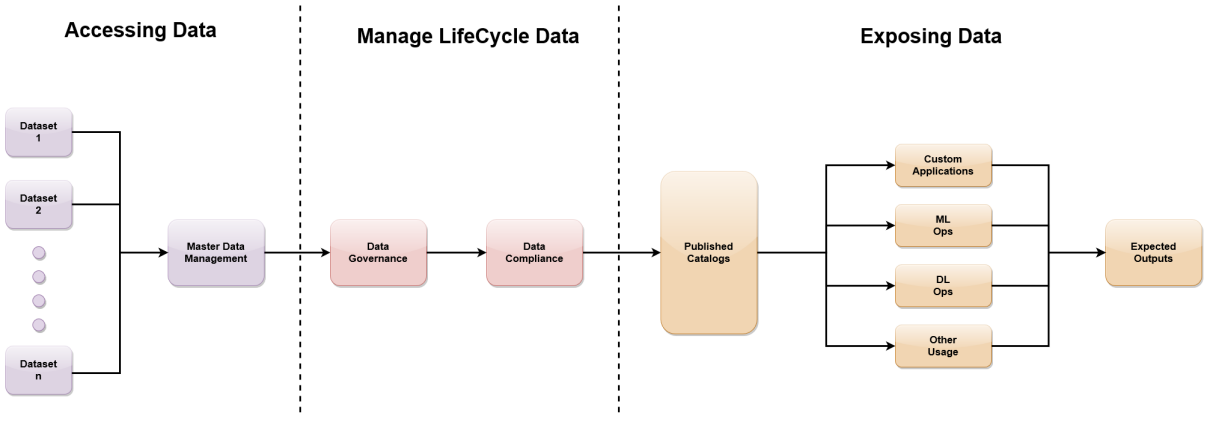


Figure 2.1: Vanilla Architecture of Data Fabric

2.2 Federated Learning

Federated Learning is a distributed machine learning technique that enables multiple clients to collaboratively learn a shared model without exchanging their raw data. This technique has gained popularity in recent years due to its ability to preserve data privacy and security while improving model performance. In Federated Learning, each client trains a local model using its own data and then sends the local model weights to a central server. The central server then aggregates the local model weights to update a global model that is shared among all clients. This process continues iteratively until the global model achieves the desired level of accuracy. According to the report [6], Federated Learning has been successfully applied to various domains, such as speech recognition, natural language processing, and healthcare, where data privacy is a major concern.

Algorithm 1 Federated Process

```
Initialize global model
Split data into shards and distribute among clients
for i in range(num_communication_rounds) do
    Initialize local model
    Send local model to each client
    for client in clients_data do
        client.model.train(client.data)
    end for
    Clients send updates to the global model
    Scale updates by the weight scaling factor
    Aggregate updates using FedMax
    global_model.update_weights(max(updates) × weight_scaling_factors)
end for
```

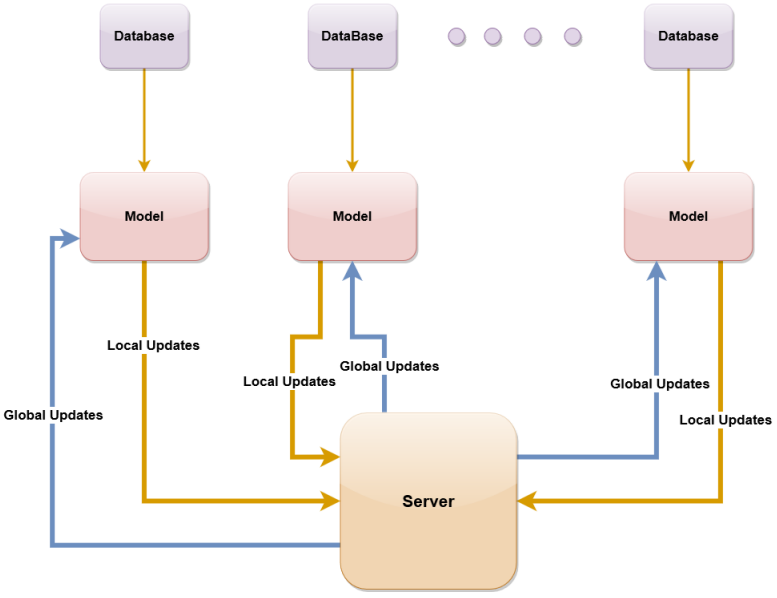


Figure 2.2: Federated Learning Model

2.3 Homomorphic Encryption

A cryptographic method called homomorphic encryption enables mathematical operations to be carried out on ciphertext without exposing the underlying plaintext. In our research, we used partially homomorphic encryption to encrypt sensitive medical images, specifically brain MRI scans.

Partially homomorphic encryption (PHE) is a type of homomorphic encryption that only supports a limited set of mathematical operations, such as addition or multiplication. By encrypting the medical images using this technique, we were able to process and analyze the data without exposing the sensitive information contained within it.

One major benefit of using partially homomorphic encryption in this context is that it ensures the confidentiality of medical data. As medical information is often highly sensitive and personal, it is important to protect it from unauthorized access. By encrypting the data, we were able to securely process and analyze it without compromising its confidentiality.

In addition, partially homomorphic encryption allows for more efficient processing of the encrypted data. Because the mathematical operations can be performed directly on the ciphertext, there is no need to decrypt the data first, which can be a time-consuming process. This was particularly useful when working with large datasets or when processing data in real time. [7]

Overall, our use of partially homomorphic encryption proved to be a successful and effective method for protecting the confidentiality of sensitive medical images while still enabling their analysis.

Types of Homomorphic Encryption	***Partially Homomorphic Encryption (PHE)***	***Somewhat Homomorphic Encryption (SHE)***	***Fully Homomorphic Encryption (FHE)***
Supported Operations	Addition or Multiplication	Addition & Multiplication	Arbitrary Computations
Security	High	Moderate to High	High
Computational Efficiency	High	Moderate	Low
Computational Intensity	Low	Moderate	High
Encryption	Simple	More Complex than PHE	Very Complex
Encryption Overhead	Low	Moderate	High
Implementation Ability	Easy	Moderate	Difficult

Table 2.1: Comparison between the types of Homomorphic Encryption

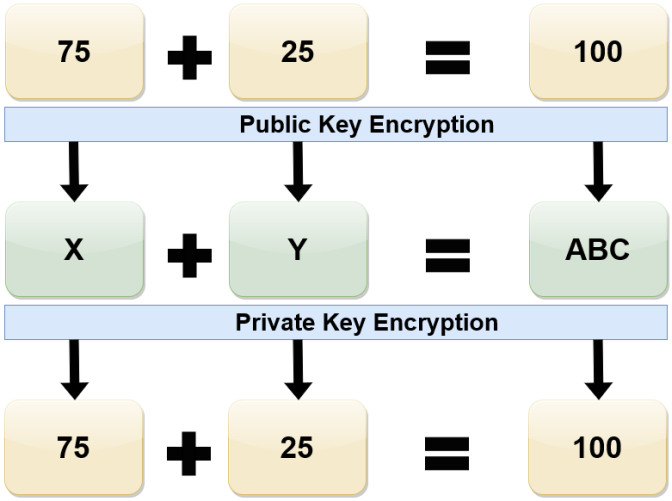


Figure 2.3: Homomorphic Encryption

Chapter 3

Related Works

Data fabric architecture is a relatively new concept that has already been utilized by notable organizations, including IBM, for practical purposes. Despite its potential benefits, there is limited research available on the implementation of this architecture in the healthcare system. Due to the sensitive nature of healthcare, developing a secure data fabric architecture can present challenges. This chapter examines various data architectures, processes, and encryption methods that can be employed to ensure the security of healthcare data.

There are a few works that are related to the architecture we are working on. They are described below:

In this paper [8], the authors have proposed a model to manage data in traffic and road systems which is based on data fabric architecture. Data will be collected on traffic events and will be kept following the principles of Blockchain networks. A transport modeling tool is planned to use here. They use Data Fabric solutions of IBM, Cloudera, and Microsoft platforms. The blockchain has been implemented using the Hyperledger Fabric framework. They have used the UNSW NB15 dataset to demonstrate the process. The authors have claimed that in this particular sector, Data Fabric performs better than Data Virtualization and is also less complex. Though they have used different types of data, they have proven data fabric can be used in real-world scenarios. However, traffic data is not as sensitive as medical data. Therefore, the exact architecture may not be suitable for medical purposes.

Due to the loss of Edge Computing opportunities because of cumbersome data assets-agnostic processes, the authors have written [9] about how data fabric driven by AI can come into use in Cloud-to-Edge Intelligence. They have created MEDAL, an efficient Cloud-to-Edge data fabric, to help DataOps automate management and orchestration processes across the spectrum using an integrated set of data and the service layer.

The authors [10] describe a proof-of-concept implementation that uses the Hyperledger Fabric framework. They claim that this concept is capable of storing patient records effectively with keeping all the privacy protocols intact. Lastly, they compared the read-write times according to their claim.

In their study, Roehrs et al. [11] divided personal health record (PHR) information into data blocks that may appear to be centrally stored but are actually distributed among participating devices. The authors claimed that their proposed openPHR protocol is practical, flexible, and scalable for adoption by multiple organizations. Although the authors provided a detailed architecture, questions have been raised about the feasibility of their approach, especially regarding security and privacy concerns. It is important to note that PHRs are controlled by patients, while electronic health records (EHRs) are managed by healthcare institutions. Nevertheless, EHRs and PHRs are electronically stored and distributed and can be assessed based on metrics such as performance, scalability, privacy protection, and compliance with the GDPR.

The MeDShare platform [12] shares several similarities with PREHEALTH [10]. However, the authors did not explicitly specify the underlying blockchain framework. Additionally, their emphasis was more on examining the fundamental components of blockchain technology, including data blocks and smart contracts, rather than presenting a practical solution.

Ming and Zhang [13] presented an effective privacy-preserving access control (PPAC) strategy for cloud-based EHR systems. Their approach utilizes the cuckoo filter and an innovative attribute-based signcryption (ABSC) mechanism to achieve both anonymity and computational efficiency. The authors offered comprehensive assurances of privacy and conducted thorough performance evaluations for comparison. However, it is uncertain whether their approach complies with the GDPR regulations.

Sharing data among different participants safely is the main focus of this paper. Using Hyperledger Fabric they have proposed a more secure decentralized distributed data storage over traditional centralized data for SKA data because of high management cost and low credible traceability. SKA Data Management Alliance has cut a significant amount of cost and improved the overall security of its data by adopting this distributed storage system [14].

In this article, [15] the authors described such an architecture that will gather real-time data focusing on the (USL) Ubiquitous Street Light to create an urban data network. Street furniture containing sensors and information display devices will be the key component in simulating the urban space into a media fabric containing data. The method Urban Media Fabric Simulator used CityGML, Location Based XML Parser, and Pattern Generation to justify the process. Because of its real-time advantages, Urban Media Fabric Simulator will massively help control media and information inference.

In our research, homomorphic encryption plays a critical role, and we have incorporated it into our architecture. We have conducted extensive work on this type of encryption, and here are some of the related studies that we have reviewed:

With the world stepping into modern days every day, privacy is becoming much of a concern; the authors [16] have used Homomorphic Encryption in Distributed

Systems to preserve the privacy of data. Data can be accessed by unauthorized people if the data on the cloud is not encrypted. The authors in this paper have used Elliptic Curve Cryptography (ECC) for the homomorphic encryption of data.

The authors [17] have provided a thorough and instructive comprehensive analysis of homomorphic encryption and its use in the health sector. Oftentimes the patients' data in the healthcare system is lost and that is why homomorphic encryption has been used so that it can protect sensitive information such that only providers can access only encrypted data by permitting data to be handled in an encrypted manner.

Authors in this article [18] decided to encode a batch of quantized gradients instead of encrypting individual gradients and did it one go by turning it into a long integer. They developed new quantization and encoding schemes along with a gradient clipping algorithm to allow gradient-wise aggregation to be performed on ciphertexts. Their MPC let multiple parties collaboratively compute with private data that did not let one party know about another except for its input and output. They managed to reach a peak accuracy of 88.62% for FE MNIST.

Our architecture utilizes federated learning to train deep learning models. In our research, we have come across some related studies that are somewhat aligned with our work. Here are brief descriptions of these studies:

In this paper [19], authors have proposed a secured model of federated learning using homomorphic encryption and tried to classify Covid 19 using X-Ray images. They have secured the federated process. They claimed that sensitive information can get into the hands of attackers if the DL process is not secured. They did not encrypt the dataset but encrypted the weight matrix that federated learning use. They claimed to obtain 84.00% accuracy with 86.89% precision.

With the growing use of homomorphic encryption, the authors have directly performed arithmetic operations on ciphertexts without decryption to protect the model parameters[20]. They decided to use a horizontal FL scenario that operates in a cloud system using SMC. Their main goal is to protect data privacy by encrypting the local model parameters using the DHC. They considered many attack scenarios in their proposed system to make sure of data privacy. The authors have used 105,506 convolutional neural networks for the simulation study. The proposed system has proved to be 2.3 times greater than that of the PPDL system.

The authors in this article [21] combined Homomorphic Encryption and Verifiable Computing techniques to produce formal proof that the operator was correctly applied and performed a Federated Averaging operator directly in the encrypted domain. A huge quantity of optimizations on the larger deep learning models was introduced to make sure practical execution performances were reached. Through this project, the authors provided an extensive experimental result on the FE MNIST dataset.

It is obvious that with the vast growth of artificial intelligence and big data, contra-

dictions between user data policy and data policy also grew proportionally. That is why the authors in the article [22] came up with a vertical federated learning system for Bayesian machine learning using homomorphic encryption. Their model could be compared with 90% models which were trained by a single union server. Additionally, their system can be used in education, financial, medical, risk controls, and other fields.

Here are some other studies that are relevant to our research and have been reviewed and summarized below:

The research work [23] provides insight into the difficulties of medical data analysis and security and proposes a solution based on a decentralized architecture. They use the Exonum framework. In their proposed architecture they have separated the whole system into two parts - 'Closed Information' and 'Open information' wherein the closed part encrypted data is stored in a blockchain and the open part non-encrypted service-related data is stored.

In this paper, [24] the role of big data in healthcare data has been explained by the authors. Their focus has been on security, privacy, and encryption. They have proposed to use hashing techniques like SHA-256 and Kerberos mechanism and encryption algorithms like SA, Rijndael, AES, etc.

The authors [25] in this article have proposed a meaningful usage of optimizing Electronic Health Records (EHRs) using big data analytics. Here, the authors proposed an insight into how to improve electronic health records using three methods: Data Collection, Data Storage, and Data Utilization. Firstly, in the data collection method, records should be divided into structured and unstructured data. Structured data may include demographics, health status, lab results, billing, etc. while surgical videos or diagnosis notes will fall under unstructured data. After collecting, they proposed a transformation engine where data will be moved, cleaned, merged, and validated and will be stored in DBMS or Cloud, or NoSQL. Finally, transformed data will be processed using mapping and reduction, and stream computing and in-database analytics will be used for generating reporting systems, which will help to achieve a meaningful usage of EHRs using big data analytics. However, the authors also mapped out the limitations of this research, and that this will lay the foundation for interesting opportunities in the future.

This research article [26] is based on the healthcare system, which the authors have reduced down to skin illnesses and have come up with solutions using an Imtidad reference architecture. This reference architecture offers Distributed AI over the cloud, as well as a service catalog case study with 22 AI skin infection diagnosis services. Throughout the study, the services have been developed using a variety of benchmarks such as the model's overall service quality, energy consumption, response time, and many more.

With the growth of Distributed AI in healthcare systems in the modern world, the authors [27] have successfully brought up research on the ethics of AI in healthcare. The authors have provided a whole mapping review on it by yielding a total of 156

papers that were included in the review. They have carried out a series of screening stages which included major descriptive data sharing, data sharing, data access, ownership of health data, consent, and evidence of efficacy.

The authors [28] have discussed in this paper how AI models are traditionally centralized in development, training, testing, etc., and are associated with such a centralized model. They have also discussed how such problems can be broken down into subproblems with decentralized data. The analysis of the Decentralized AI has proven that it's in great growth and functioning in the field.

Since modern organizations have a universal need for AI services, the authors [29] have presented an effective path for AI development that is a decentralized, and democratized market for AIs to work on distributed ledger systems. They created a system called SingularityNET to define the qualities and ethical benefits of watershed initiatives in development.

In this article, [30] the authors described a linear time-distributed algorithm for the state of being between two others in an ordered mathematical set of computing centralities. The algorithm counts the relative validation and importance of a node within a network. It is constructed following the CONGEST model along with a lower bound therefore it limits the message to only containing O(log N) bits. Micro parts of the algorithm are built with BFS and DFS to determine the diameter of a graph along with the BC of each node. The algorithm is a distributed version of the prominent Brandes algorithm and only works on unweighted graphs.

In this article, [31] the authors described an ADS infrastructure which is a software-based DF composed of Nodes. The ADS infrastructure identifies entities and redirects messages using a communication network where each node is a software application with certain communication protocols. The article also provides a secure DF connection for identifying connected components with the help of Public Key Cryptography. The identification protocol and the cryptographic treatment for this communication protocol.

In this paper [32], the authors attempted to describe a big data platform that will be able to efficiently and effectively process healthcare data. They proposed a system that will be lightning-fast, will support stream processing, and will have integration with both NoSQL and RDBMS. This process aims to exploit open-source technologies as much as possible and build the system on top of that. The core of the system will be Spark core, and to utilize that the system will use the Spark framework with real-time graphical image processing. For handling structured data, the authors proposed SparkSQL Structured Data and MLib Machine Learning for classifying the data. This paper proposes a novel big data platform that will be able to redesign modern medical data and thus bring an effective and quick solution to the healthcare system.

This paper [33] provides a strategic planning reference to help prospective customers embarking on cloud projects. Similarly, the HC^2SP model is recommended. This can be used by healthcare institutions to figure out the direction, technology, and

distribution of resources for the cloud paradigm transition. The model contains four phases: distinguishing proof, assessment, activity, and follow-up. The organization assesses the existing situation of the service operation and identifies the basic service objectives in the first stage. According to the author, this should entail at least the following: determining cloud services and delivery methods, considering potential cloud providers, obtaining validation of approved cloud vendors, considering future data moves, and initiating experimental implementations. The next stage is to disseminate the cloud computing infrastructure and create a follow-up strategy to track how well healthcare services are improving.

In this article, [34] the authors described an architecture where an improvised big data model is involved to create a cloud computing environment for healthcare. In this process, huge amounts of data from medical sources and processes are collected in cloud storage, and real-time analysis is done using cloud computing for better accuracy. The Healthcare Data Management Framework is built on Hadoop Clusters and certain key components. Semantic Practitioner, Big data container and processing layer, Query formulator, Batch scheduling, and Data reader are examples of such tools. This architecture is also open to implementing various cryptographic techniques on Cloud.

Cloud computing has newly emerged as a new sculpt for conveying and facilitating Information Technology (IT) services over the web. In this article [35], the authors talked about different types of service models and deployment models of cloud computing such as SaaS, PaaS, IaaS and public cloud, private cloud, community cloud, and hybrid cloud. The paper highlighted how the current trend of adopting cloud computing in the clinical field can improve and settle a few cooperative data issues in medical care associations as well as cost advancements. The authors also mentioned some difficulties, for example, security concerns and interoperability will increase because of the cloud computing model. The execution of best practices in the plan, organization, and utilization of it will ideally produce a future development of the cloud-based framework's reception, despite all of the impediments.

In this paper [36], the authors proposed a cloud computing-based Administration Oriented Architecture (SOA) to handle the evolution and merging of heterogeneous datasets and application frameworks. When a research group needs access to a remote resource (data or application), it simply requests the web service through SOA by applying a normalized programming grammar. They used a cloud algorithm to virtualize the server using Virtual Machine ware (VMware) on the platform. This allows analysts to use VMware to open virtualized servers and do the same research in different areas. Various site specialists can use virtualized servers to share and modify findings while storing their patient information in nearby datasets. In addition, they used a technique called Ambiguity to ensure both the presence of privacy and affiliation security with little loss of information. The advantage of this design is that it protects the security and privacy of the original data.

In this paper [37], the authors hoped to gain insight into the concept of cloud computing for IT healthcare research. They obtained information about cloud computing services in healthcare organizations and proposed a scientific classification (taxon-

omy) that acts as a baseline system for a clear understanding of healthcare cloud computing. The scientific classification focused on conceptualizing the applicable properties of cloud computing for servicing healthcare organizations and showing their specific impact on healthcare delivery. They used two approaches to facilitate the scientific classification of cloud computing services for healthcare organizations. By applying scientific classifications to group existing cloud computing services that differ from literature and expert interviews (part of the scientific classification), they recognized a few features of cloud computing in health care.

The authors of this research [38], proposed a security paradigm for maintaining the confidentiality of clinical data in the healthcare cloud by combining a cloud computing capability with encryption. In this journal, a secure authentication mechanism based on piecewise linear pairing cryptography is proposed, which will be capable of generating and securely distributing a session key among healthcare cloud participants. After distribution, a decoy technique will help to safely store data and securely access it.

The writers in [39] expressed their concern over security and privacy in cloud computing platforms and protecting data "in processing". For this reason, they created a trust-based paradigm for MapReduce Big Data Processing. A MapReduce scheduling problem is implemented here, with a bipartite graph determining the total reliability of all feasible task assignments. Each task receives a specific amount of confidence based on the sensitivity level of the data, and thus a heuristic-based algorithm will find out the optimal solution and provide higher protection for data sensitivity.

Chapter 4

Methodology

4.1 Data Description

For performing operations, we collected the dataset [40] from Kaggle named "Brain MRI Images for Brain Tumor Detection". The dataset includes a large number of 2D MRI images for the classification of brain tumors. Brain tumors are classified into three types: Benign, Malignant, and Pituitary. However, for our selected dataset, images are classified into two categories; Pituitary Tumor and No Tumor.

From the dataset, 1852 images have been used by us as shown in the figure below. On the images, we performed encryption and machine learning techniques for ensuring privacy and achieving desired results.

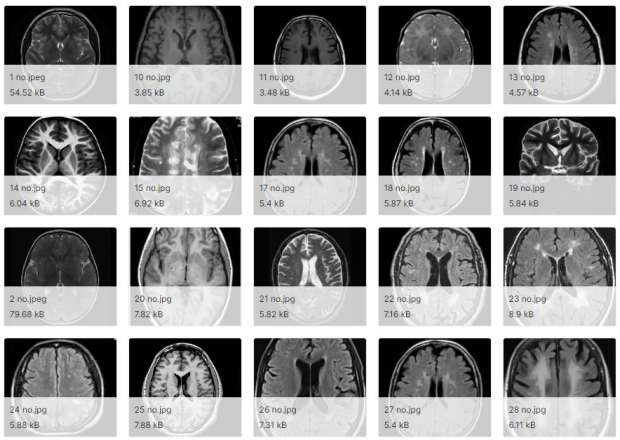


Figure 4.1: Collected dataset from Kaggle [37]

For preparing our own data fabric to perform MLOps, an image was selected from the dataset and converted into a Numpy array with its pixel values. After that, homomorphic encryption was performed on the array to make the data fabric encrypted. A brief overview is shown in the following figure.

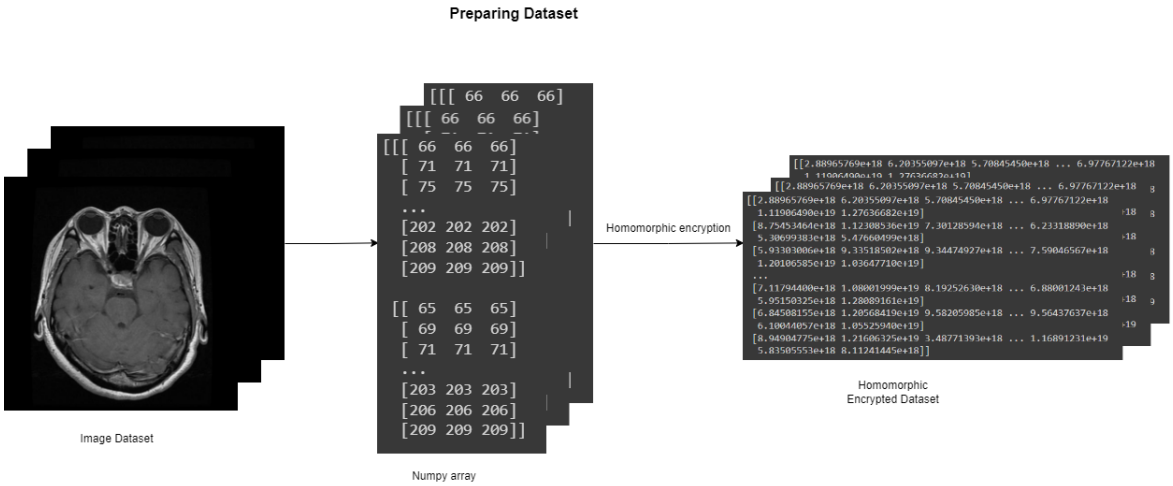


Figure 4.2: Overview of the encrypted dataset

4.2 Preprocessing Data

To prepare our data we needed to preprocess the whole dataset. Firstly, we have encrypted our dataset using the Partially Homomorphic Encryption Algorithm. After that, we resized the images to 128*128 dimensions. The shape of our dataset became (1852, 128, 128). Then we reshaped our dataset by multiplying the dimensions of individual pixels. Then the shape ultimately (1852, 15376). For machine learning classifiers we have scaled the image dataset to value 0 to 1.

4.3 Proposed Model

4.3.1 Advanced Architecture of Data Fabric

1. **Accessing Data:** Here, initially, to ensure privacy and reduce data volatility, medical data from various users are first selected and then encrypted with Partially Homomorphic Encryption (PHE). Partially Homomorphic Encryption (PHE) enabled us to perform DLOps on the data securely. Subsequently, the encrypted data is selected to train the model locally for collecting updated model weights. For training, various federated learning models like VGG16, VGG19, ResNet50, ResNet152, and our Custom CNN were used which generated model updates for each model. The data of each user is generated and stored locally, without being transferred to the central server. Instead, the local model updates were stored and merged for the global model formation.

   After combining the local model weights, we selected FedMax as our feature selection algorithm to select important and relevant model weights as it showed more accuracy and efficiency compared to FedAvg and FedMin. The selected data is collected and kept together as “Master Data”, with which we will continue our next works.

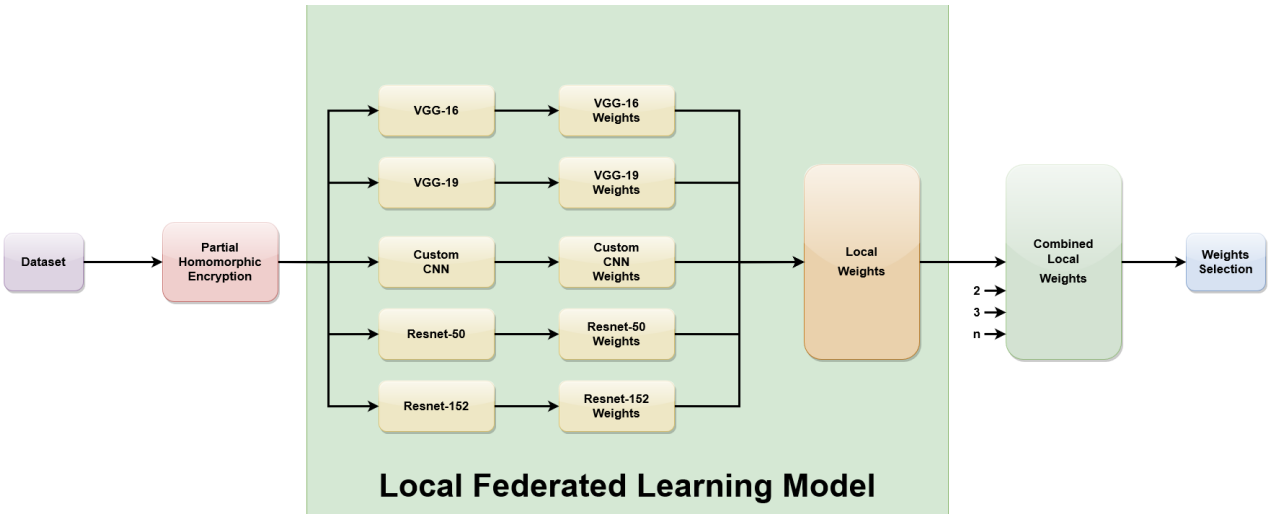


Figure 4.3: Accessing Data

2. **Managing Life Cycle:** Selected model weights were differentiated based on the models they were trained on. In this case, for example, updates of the VGG16 model were kept structurally under the "VGG16" name. This enables effective data governance, as the risks of data inconsistency and complicated integration across the whole architecture get lowered. Additionally, organizing data in a structured way ensures proper usage of data and helps strike a balance between data collaboration and privacy mandates.

   Furthermore, as the privacy of the data was already ensured in the first step, the collected model weights already comply with existing privacy regulations such as HIPAA, GDPR, etc.

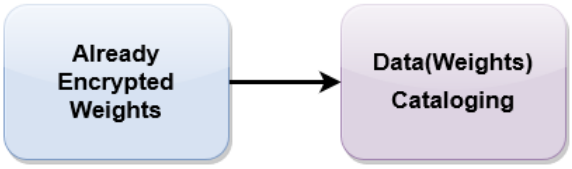


Figure 4.4: Data Lifecycle Management

3. **Exposing Data:** Since exposing data is a comprehensive approach where appropriate data access controls, integration, and cataloging must be implemented, we stored all the collected local model updates in a "Data Lake". This enabled us to store our raw model weights structurally in their native format. Additionally, since there were huge amounts of model weights, using a Data Lake helped us to store and process it easily.

   Most importantly, we can also perform MLOps on the Data Lake directly. On the client side, clients can train their data on the global federated model and

can compare the global weights with the local weights of that model stored in the data lake and generate the desired output.

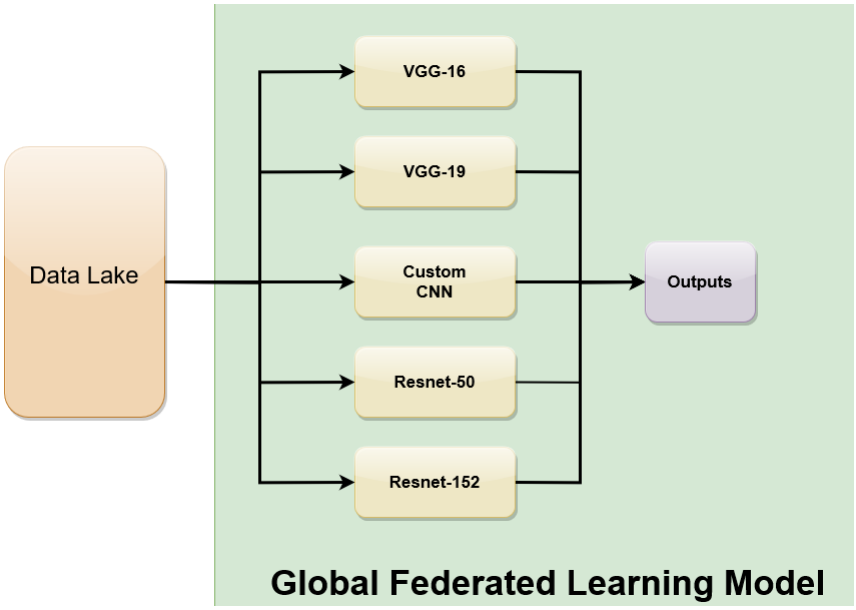


Figure 4.5: Exposing Data

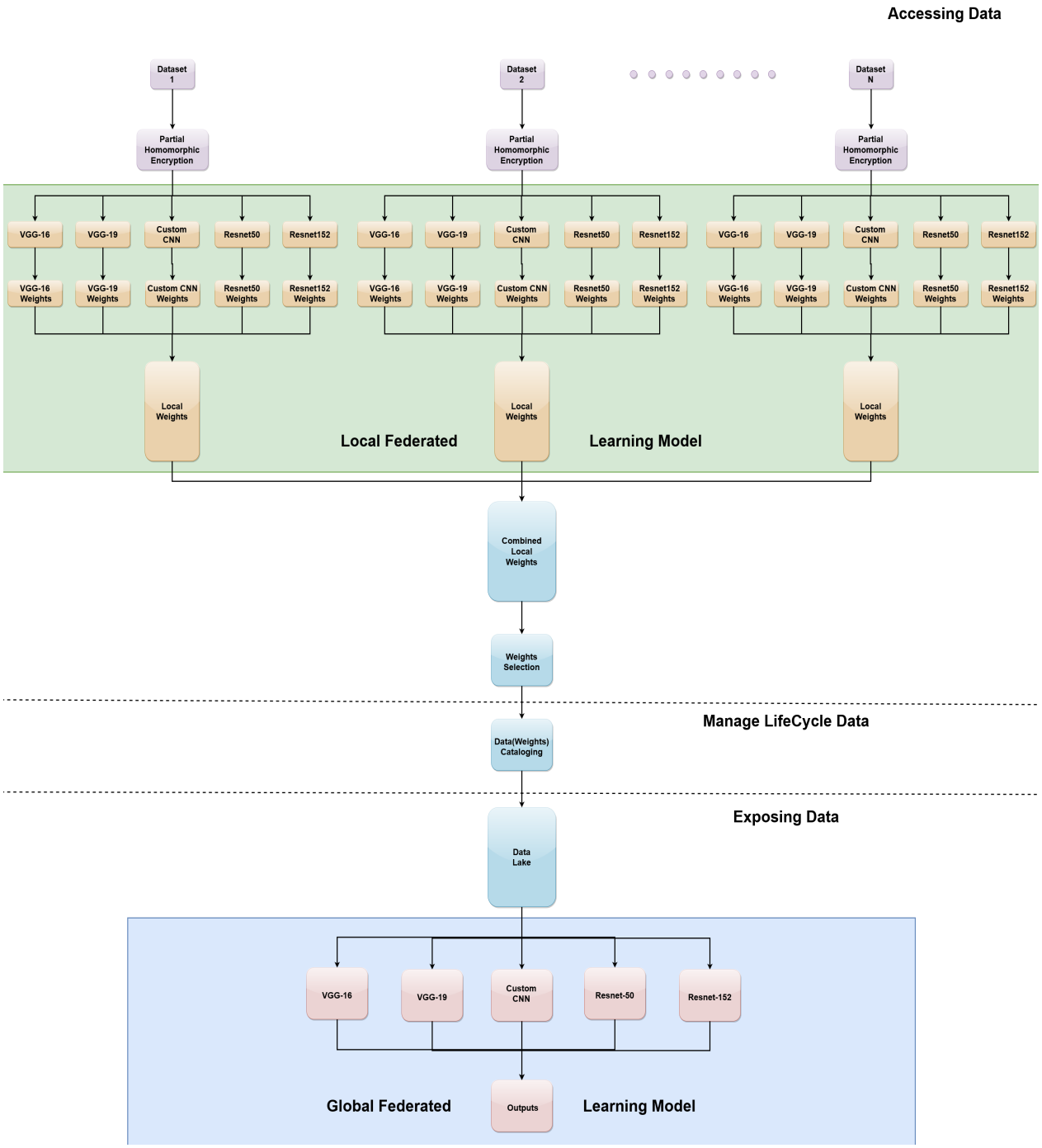


Figure 4.6: Proposed Architecture of Data Fabric

4.3.2 Implemented Federated Learning Framework

As we delve into the world of medical imaging and machine learning, we have utilized a cutting-edge approach to store the local weights of our machine learning models using a Data Fabric architecture. This architecture has allowed us to securely store and manage distributed data across multiple environments, providing a consistent and scalable approach to managing data assets.

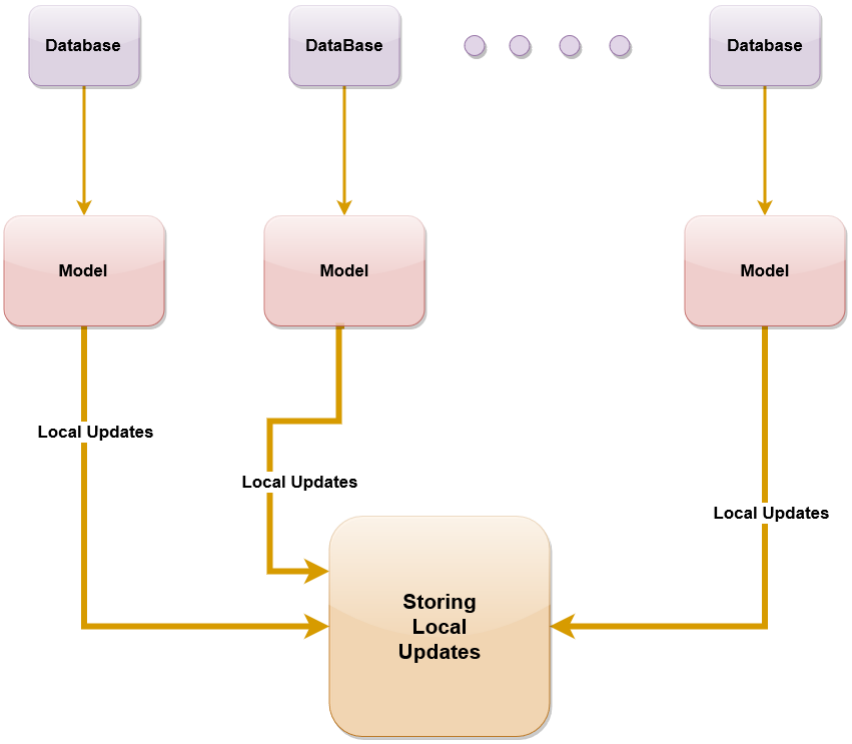


Figure 4.7: Local Federated Learning Model

Our goal was to develop a pituitary tumor classification model that leverages the decentralized data in a privacy-preserving manner, improving the accuracy of the model while ensuring the privacy of patient information. We have used the FedMax algorithm, an improvement on the popular FedAvg algorithm, to compile a global model and local models for the classification of pituitary tumors from MRI images in our dataset. The FedMax algorithm considers the model performance of each client and assigns a weighting factor that allows clients with the best performance to contribute more to the global model. We have stored the resulting model weights in a data lake, allowing us to generate a global model and test it using test cases when the user prompts to show results.

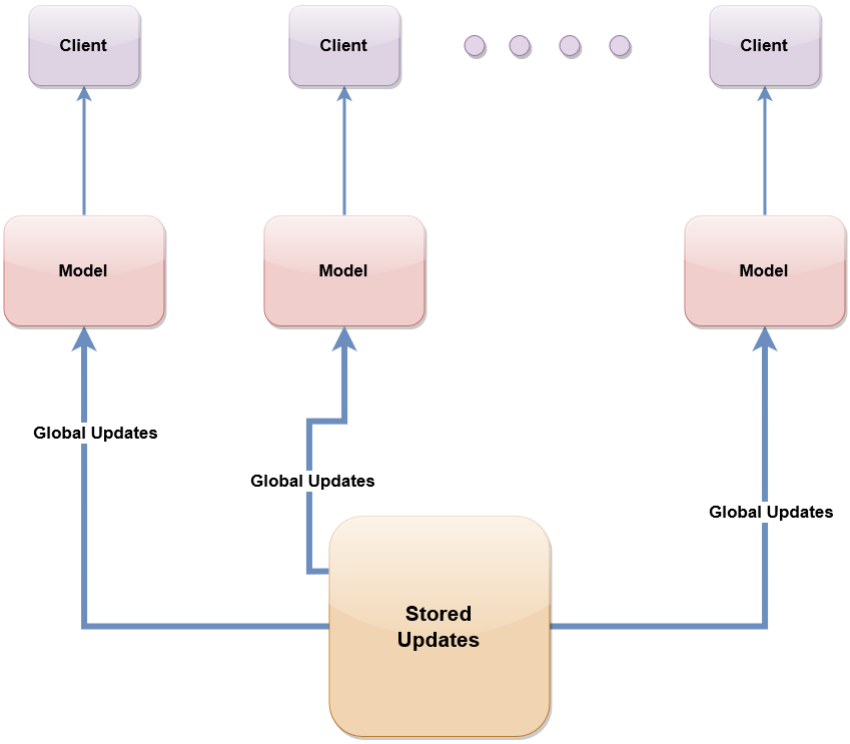


Figure 4.8: Global Federated Learning Model

4.3.3 Comparative Analysis with Related Works

Our work is different from existing approaches as it enables effective and secure handling of highly sensitive health data through our proposed Data Fabric architecture. Another advantage of this work is that its privacy-preserving features (Homormophic encryption) are compliant with GDPR [41], as it supports the right to be forgotten, as the actual health data is neither collected nor stored. Rather, after training, the encrypted data which is collected is stored as model updates. Similarly, it is possible to delete a client's data from the data lake upon request. In addition to this, compared to other works, our model also complies with HIPAA guidelines, as it ensures confidentiality, integrity, and availability of personal health information, safeguards data from threats, and protects impermissible access as mentioned here[1].

The following table displays a comparison between our data fabric architecture and other existing works that have demonstrated proof-of-concept implementations, which can be implemented and practical for real-world scenarios. In this table, the technology, privacy-preserving features, GDPR and HIPAA compliance, and also performance assessment results of each existing proposal have been provided.

Architecture	Technology	Access Level	HIPAA	GDPR	Privacy Preserving	Performance
PREHEALTH [10]	Hyperledger Fabric	Private	Not Mentioned	Yes	Yes	Yes
OmniPHR [11]	Peer - to - Peer	Private	No	No	No	Yes
MeDShare [12]	Agnostic	Open	No	No	No	Yes
Access Control Based EHR [13]	AC Scheme	Private	No	No	Yes	Yes
Our Proposed Work	Data Fabric	Private	Yes	Yes	Yes	Yes

Table 4.1: Comparative Analysis

4.3.4 Model Specification

1. **VGG16:** A 16-layered convolutional neural network model trained on the ImageNet dataset, it is considered to be one of the best models to date. It is widely regarded for its simple architecture and excellent image classification performance, retaining 92.7% test accuracy in the ImageNet dataset, which consists of almost 14 million training images across a thousand object classes.

   As its name suggests, it is composed of 16 layers which include 13 convolutional layers and 3 fully connected layers. It comprises 138 million parameters and uses small convolutional filters and deep architectures to gain a large receptive field and strong discrimination ability, which helps in image classification, object detection, and semantic segmentation. To control overfitting and reduce spatial dimensions, its architecture has five max pooling layers.

   The most unique thing about VGG16 is that it is focused on convolution layers of 3*3 filter with stride 1 and always used the same padding and max pol layer of 2*2 filter with stride 2, and the layers are constantly arranged over the whole architecture. Due to its high-level feature representation, it provides good performance on object detection, fine-grained image classification, etc.

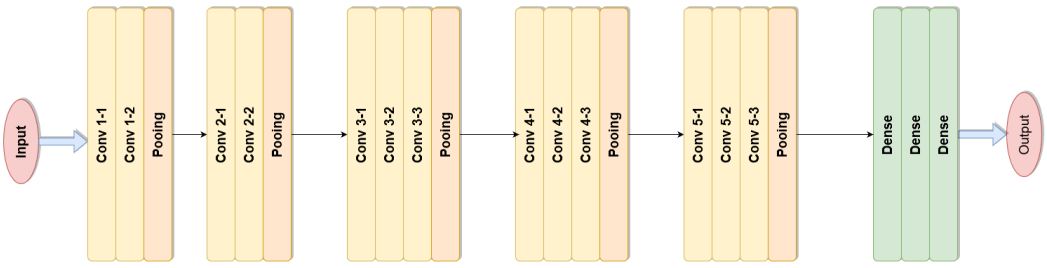


Figure 4.9: VGG16 Model Architecture

2. **VGG19:** A variant of the VGG neural network, it is a 19-layer version of the VGG network and similar to the VGG16 architecture. Compared to VGG16, it has 5 convolutional layers and 1 fully connected layer. It has around 143 million parameters and is trained on a dataset with 1.2 million images and 1000 classes.

The function accepts an image of shape (128, 128, 1) as input, and the image is passed through concatenate layer which concatenates the image 3 times resulting in a tensor with shape (128, 128, 3), this is passed to a VGG19 model which is a pre-trained convolutional neural network model that is trained on the ImageNet dataset. The VGG19 model serves as a feature extractor and it extracts features from the image by stacking convolutional layers and pooling layers. The output of the final max pooling layer is then passed through a flattened layer which reshapes the output tensor into a 2D array. This flattened layer's output is passed through a dense layer with 1 unit and a sigmoid activation function, it produces the final output of the network which represents the predicted probability of the input image belonging to the target class.

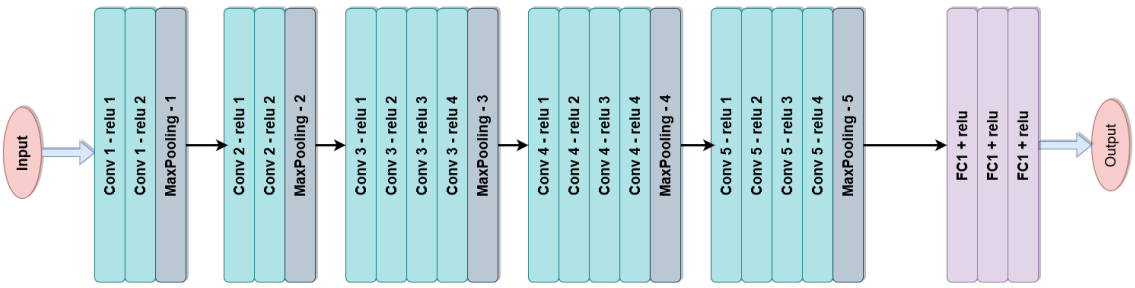


Figure 4.10: VGG19 Model Architecture

3. **ResNet50:** ResNet50 is a deep convolutional neural network architecture that is trained on more than a million images from the ImageNet dataset. It is known for its use of residual blocks, which address the problem of vanishing gradients in deep networks. These residual blocks have shortcut connections that bypass one or more layers, allowing gradients to flow more easily and making it possible to train very deep networks without the problem of vanishing gradients. The architecture also uses batch normalization layers, which normalize the activation of the layers, making it possible to train the network more quickly and effectively. Additionally, the number of filters increases as the network gets deeper, allowing it to automatically learn more complex features. ResNet50 is widely used in many computer vision tasks and is commonly used as a feature extractor for other tasks.

   The ResNet50 model is used as a feature extractor in our experiment, and the output of the final convolutional layer is passed through a flattened layer which reshapes the output tensor into a 2D array. The flattened feature map is then passed through a Dropout layer with a drop rate of 0.5, which is used for regularization to prevent overfitting. The dropout layer is optional and could be removed by commenting out the line. Finally, the output of

the dropout layer is passed through a dense layer with 1 unit and a sigmoid activation function, which is used to produce the final output of the network, which represents the predicted probability of the input image belonging to the target class.[42]

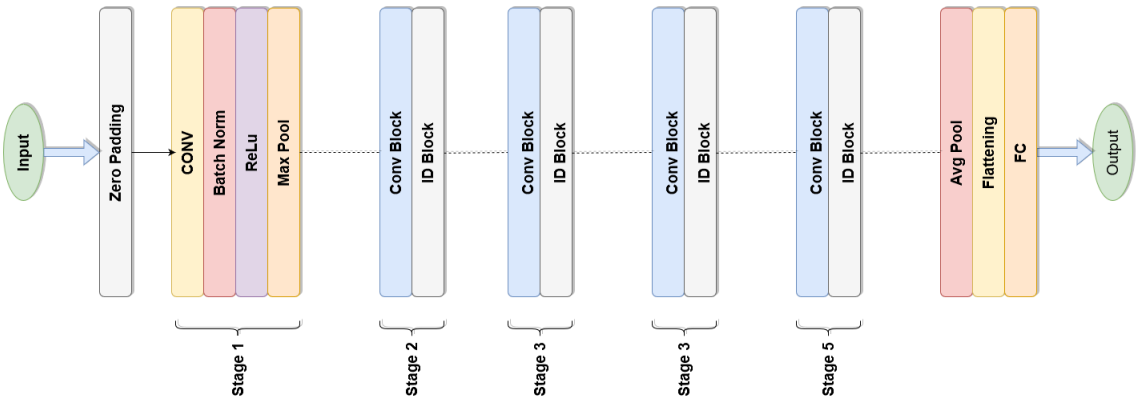


Figure 4.11: ResNet50 Model Architecture

4. **ResNet152:** We also used a ResNet152 architecture model. The architecture is composed of a stack of convolutional and pooling layers. The model accepts an image of shape (128, 128, 3) as input and the output of the final average pooling layer is passed through a Flatten layer which reshapes the output tensor into a 2D array. Then, there is a Dense layer with classes number of units, and softmax activation function. This dense layer produces the final output of the network which represents the predicted probability of the input image belonging to different classes. This architecture does not have a dropout layer and the model is trainable, this means that the model can be fine-tuned with new data for a different task. [42]

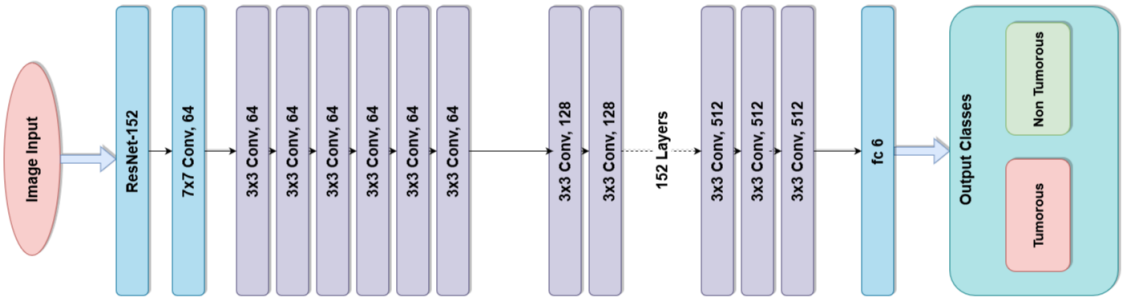


Figure 4.12: ResNet152 Model Architecture

5. **Custom CNN:** We have customized a convolutional neural network (CNN) in a similar structure to existing VGG16 and VGG19 models. It has a total of 16 convolutional layers and 3 fully connected layers. The architecture starts with

an image of size (128, 128, 1) as input and predicts a binary output. The first Conv2D layer consists of 64 filters of size 3 x 3, followed by another Conv2D layer of 64 filters of size 3 x 3, and a max pooling layer with a pool size 2 x 2. The next two layers are similar, with 128 filters of size 3 x 3 and another max pooling layer. This is followed by several more pairs of DepthwiseConv2D with a different number of filters and MaxPooling2D layers that are stacked on top of each other and learn increasingly complex features from the images. After two fully connected layers with 4096 units and a dropout of 0.5 applied on each layer, the output is generated with activation "sigmoid" and Dense Layer 1 unit. Compared to VGG16, our customized CNN has a deeper architecture with more convolutional layers and is similar to that of VGG19. It also has a higher number of filters in layers than VGG16 and VGG19. Additionally, Custom CNN uses a dropout layer after each fully connected layer, something that is not present in VGG16 and VGG19.

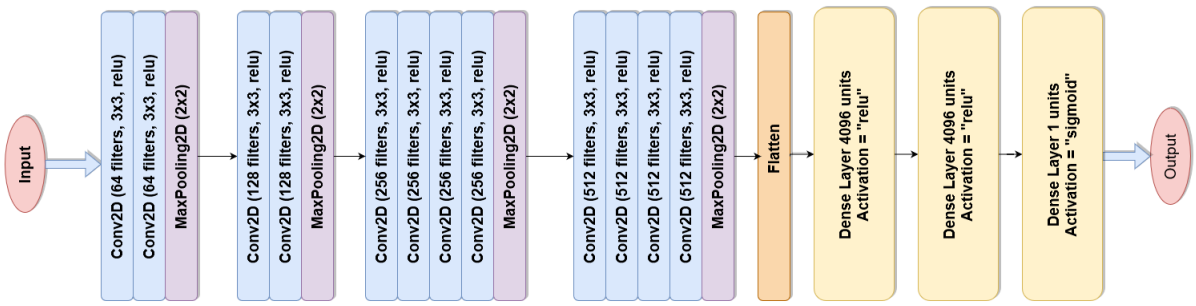


Figure 4.13: Custom CNN Model Architecture

Chapter 5

Result Analysis

We used homomorphic encrypted MRI data for our experiment. In our architecture, this part works as data governance. We tested simple machine-learning algorithms on both encrypted and unencrypted data. We saw on unencrypted data we were able to get the highest 97.1% accuracy whereas, on encrypted data, we were able to get the highest 70.12% accuracy. That shows The approach that we followed to encrypt the data is somewhat usable.

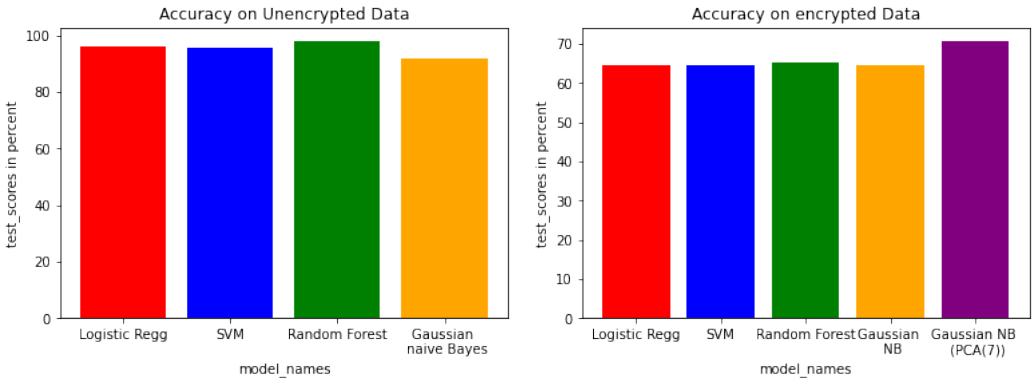


Figure 5.1: Performance of tested models on unencrypted and encrypted dataset

For our experiment, we used four different pre-trained deep learning models - VGG16, VGG19, ResNet50, and ResNet152 - as well as a custom CNN model that we developed in-house, to classify pituitary tumor and no tumor from homomorphic encrypted MRI data during federated learning. We evaluated the performance of these models using accuracy and F1-score, as well as precision for each class. These results are from global models of federated learning which we used in our architecture.

Accuracy is the proportion of correctly classified cases out of all cases. In this study, all five models achieved accuracy greater than 50%, indicating that they performed better than random guessing. The Custom CNN model achieved the highest accuracy of 83.31%, followed by VGG19 with an accuracy of 78.58%, VGG16 with an accuracy of 77.25%, ResNet152 with an accuracy of 65.09%, and ResNet50 with an accuracy of 61.51%.

F1-score is a harmonic mean of precision and recall and is often used as a measure of overall model performance. Looking at the F1-score results, we can see that the Custom CNN model achieved the highest score of 83.13%, followed by VGG19 with an F1-score of 78.31%, VGG16 with an F1-score of 77.11%, ResNet152 with an F1-score of 64.46%, and ResNet50 with an F1-score of 61.45%.

Model	Accuracy Scores		Precision		Recall	
	Accuracy	F1 - Score	Pituitary Tumor	No Tumor	Pituitary Tumor	No Tumor
VGG16	77.25%	77.11%	75.28%	79.22%	80.72%	73.49%
VGG19	78.58%	78.31%	75.82%	81.33%	83.13%	73.49%
ResNet50	61.51%	61.45%	62.38%	60.67%	57.83%	65.06%
ResNet152	65.09%	64.46%	68.18%	62.00%	54.22%	74.70%
Custom CNN	83.31%	83.13%	85.71%	80.90%	79.52%	86.75%

Table 5.1: Accuracy scores across different models

The VGG16 model had a precision of 75.28% for pituitary tumors and 79.22% for no tumors in the binary classification task and 80.72% and 73.49% recall percentages accordingly. This indicates that the model had a moderate ability to correctly identify true positive cases, with slightly higher precision for no tumor cases than for pituitary tumor cases.

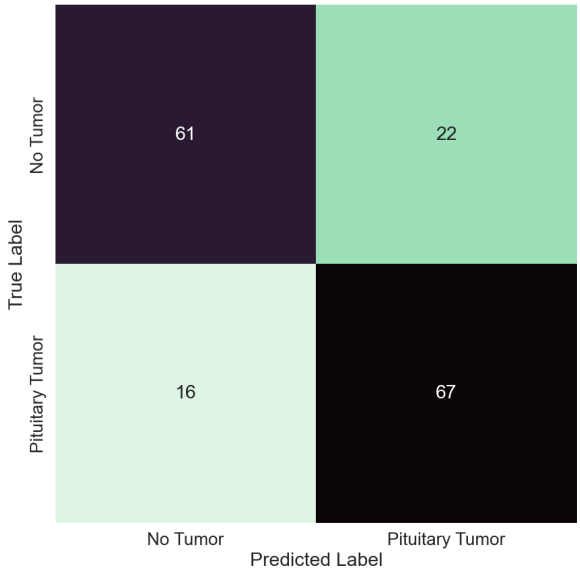


Figure 5.2: Confusion Matrix of VGG16

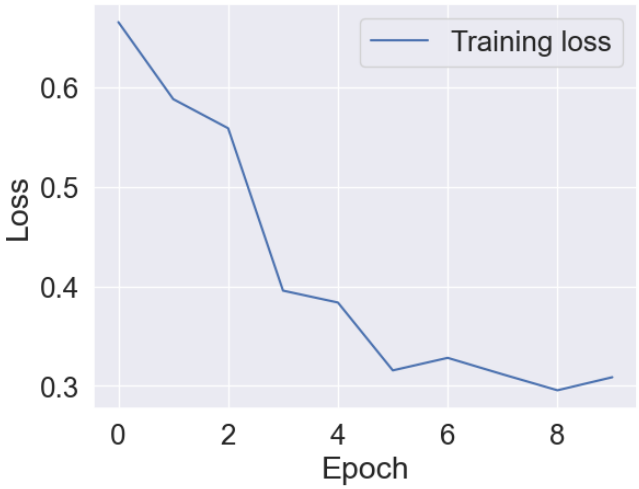


Figure 5.3: Training History of VGG16

The VGG19 model had the second-highest precision for both classes in the binary classification task. The precision for the pituitary tumor was 75.82% and for no tumor was 81.33%. The recall percentage is 83.13% and 73.49% for the classes. This suggests that the model had a moderate ability to correctly identify true positive cases, with slightly higher precision for no tumor cases than for pituitary tumor cases.

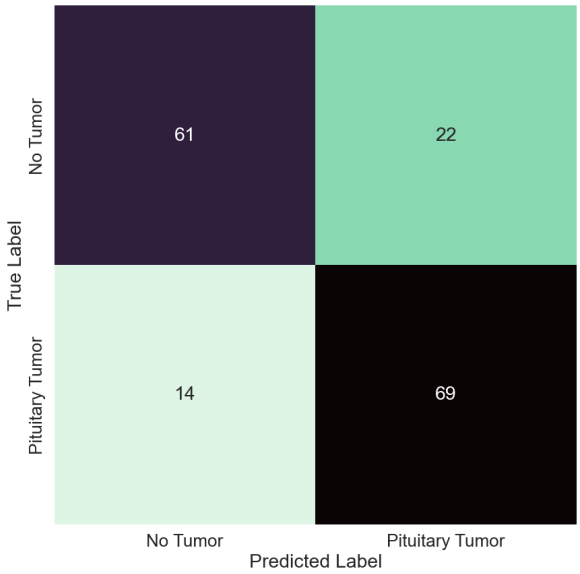


Figure 5.4: Confusion Matrix of VGG19

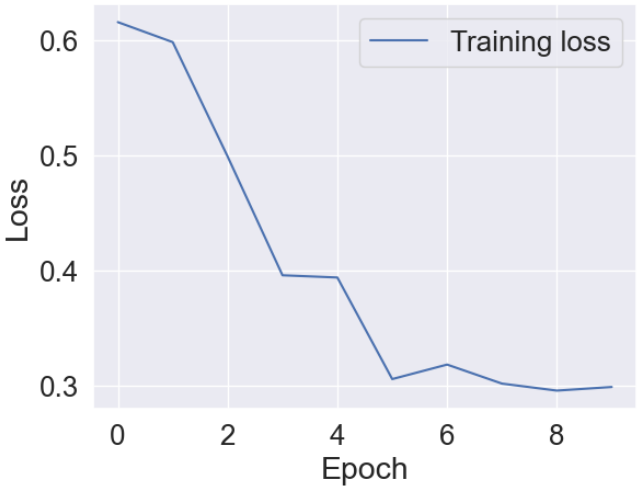


Figure 5.5: Training History of VGG19

The ResNet50 model had the lowest precision for both classes in the binary classification task. The precision for the pituitary tumor was 62.38% and for no tumor was 60.67% and the recall percentage was 57.83% and 65.06% for the classes. This indicates that the model had a lower ability to correctly identify true positive cases, with a higher number of false positives compared to the other models.

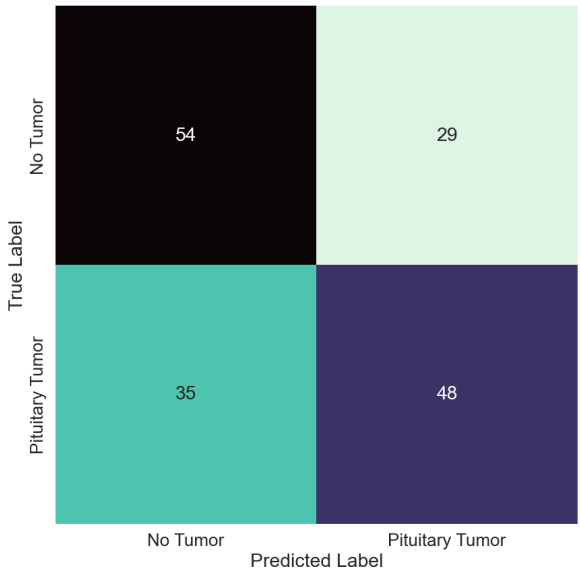


Figure 5.6: Confusion Matrix of ResNet50

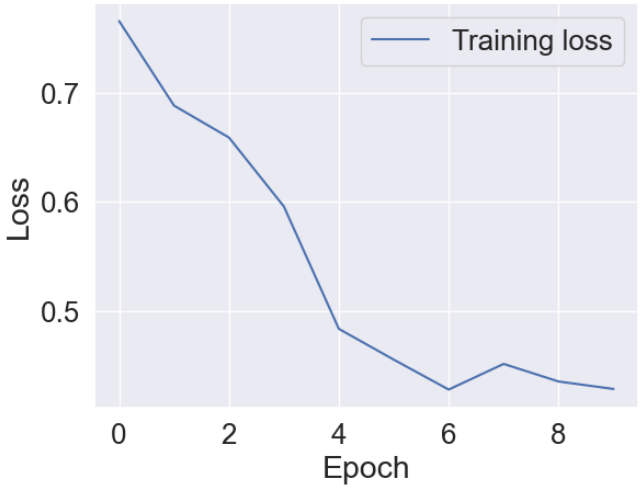


Figure 5.7: Training History of ResNet50

The ResNet152 model had a precision of 68.18% for pituitary tumors and 62.00% for no tumors in the binary classification task while having 54.22% and 74.70% recall. This suggests that the model had a lower ability to correctly identify true positive cases, with a higher number of false positives, particularly for no tumor cases.

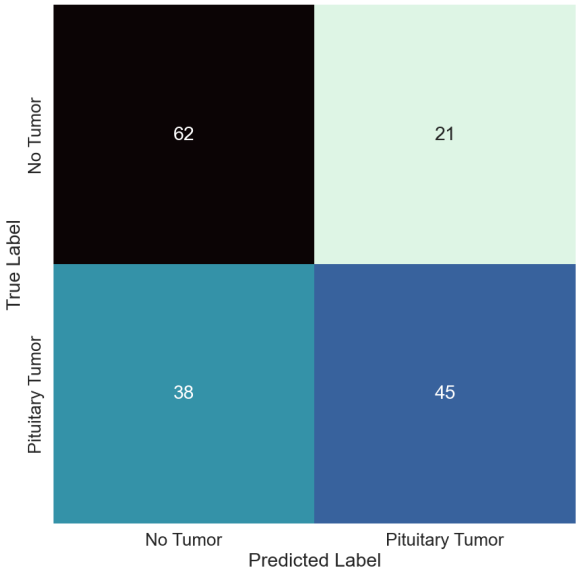


Figure 5.8: Confusion Matrix of ResNet152

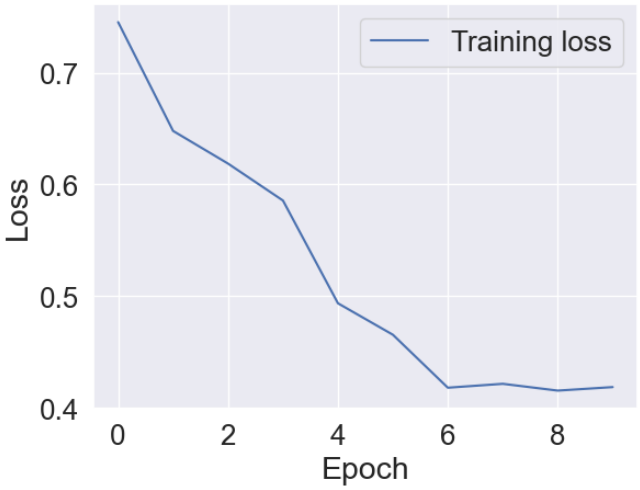


Figure 5.9: Training History of ResNet152

The Custom CNN model achieved the highest precision for both classes in the binary classification task. The precision for the pituitary tumor was 85.71% and for no tumor was 80.90%. The recall percentage is 79.52% and 86.75% accordingly. This indicates that the model had a high ability to correctly identify true positive cases, with a relatively low number of false positives.

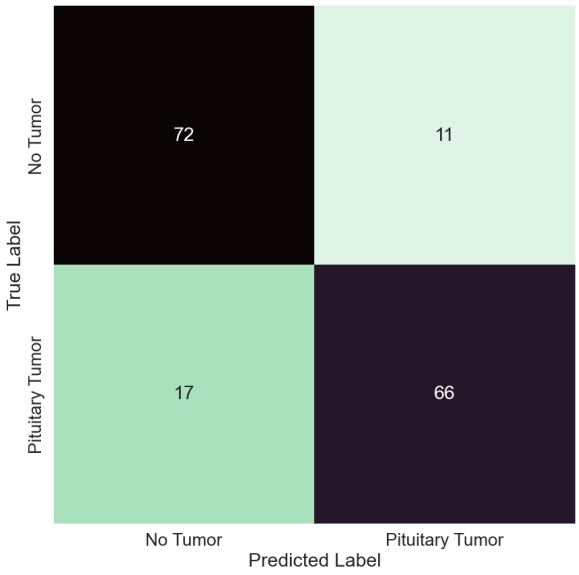


Figure 5.10: Confusion Matrix of CNN

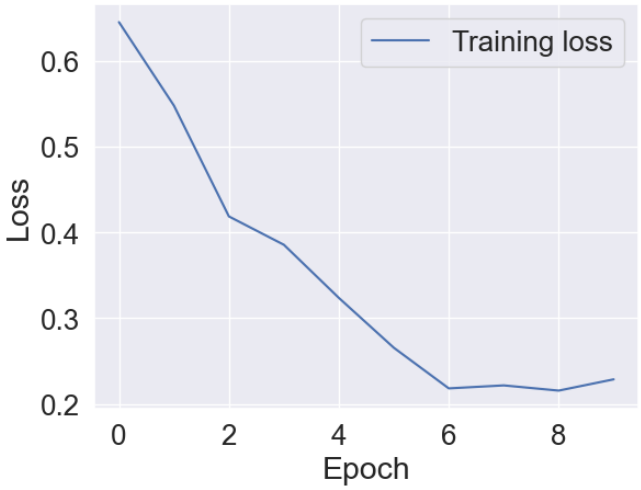


Figure 5.11: Training History of Custom CNN

The ROC (Receiver Operating Characteristic) curve is a powerful tool for evaluating the performance of binary classification models. It provides a graphical representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) at various threshold settings, allowing you to choose an appropriate threshold based on your specific needs. A good classifier should have a ROC curve that is as close as possible to the upper left corner of the graph, indicating high TPR and low FPR. By analyzing the ROC curve, you can gain insights into the strengths and weaknesses of your model and make informed decisions about how to improve its performance. The ROC curve is widely used in various fields, including medicine, finance, and machine learning, and is an essential tool for anyone working with binary classification models. The ROC curves of our used DL models are shown below:

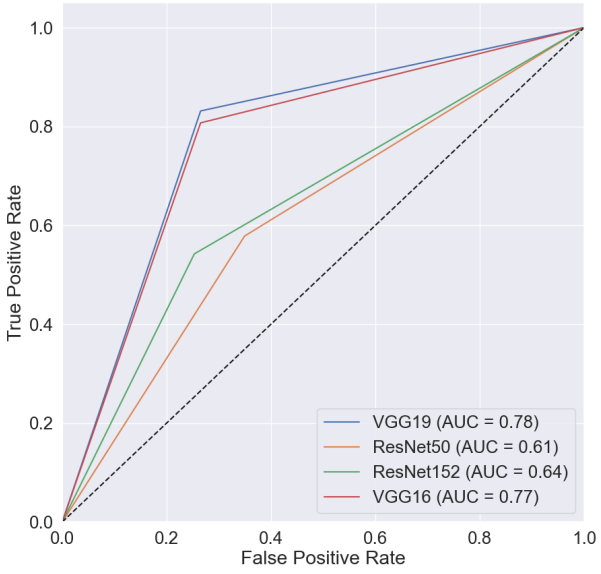


Figure 5.12: Receiver Operating Characteristic of Different Models

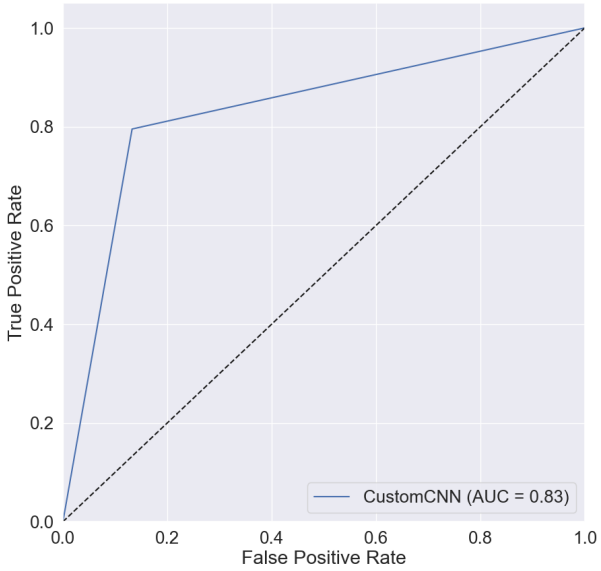


Figure 5.13: Receiver Operating Characteristic of Custom CNN

Overall, the results suggest that the Custom CNN model performed the best in classifying pituitary tumors and no tumors from homomorphic encrypted MRI data, achieving the highest accuracy, F1-score, and precision for both classes. This is particularly notable as the Custom CNN model was developed in-house, suggesting the potential for further research and development in this area. However, it is important to note that the use of partially homomorphic encryption may have influenced the results, and further research is required to confirm this hypothesis.

Chapter 6

Conclusion

Our study demonstrates an advanced data fabric architecture that ensures data security of sensitive medical images while getting the most features from these images which are needed in the advancement of artificial intelligence in the healthcare sector. We have successfully done our experiment which ensured data integrity, and privacy and also provided us with satisfactory results. We explored the use of pre-trained deep learning models to classify pituitary tumors from homomorphic encrypted MRI data. Our analysis showed that the VGG16 and VGG19 models outperformed the ResNet50 and ResNet152 models in terms of accuracy, precision, recall, and F1-score for both classes. We achieved overall satisfactory accuracy from VGG16 and VGG19. Out of all of these pre-trained models, our custom CNN model performed better with 83.31% accuracy. The reason for the model to perform better is we have made the model similar to VGG16 and VGG19 while making it much less complex. Our model has a total of around 15 million parameters compared to 138 million of VGG16. These findings suggest that the models have the potential for use in medical diagnosis and may help improve outcomes by enabling faster and more accurate diagnoses.

Despite several significant limitations, the experiment demonstrates that our advanced data fabric architecture has the potential to overcome many of the significant challenges that researchers encounter while working with sensitive medical image data. Moving forward, we can expand our research by incorporating data from all medical tests based on images and utilizing heterogeneous models to enhance diagnostic accuracy and improve outcomes.

6.1 Limitations

In our research, we encountered several limitations that hindered our ability to achieve more satisfactory results. One major limitation was the use of Partially Homomorphic Encryption, which only allowed for addition or multiplication operations. However, our data preprocessing, models, and federated process required the ability to perform both operations. This limitation greatly impacted the quality of our dataset and hindered our ability to achieve optimal results.

One of the major constraints of our research is the availability and quality of healthcare data. Healthcare datasets are not available widely and are subject to strict

privacy regulations, such as HIPAA in the United States. This limits access to the data and requires additional measures to process the data, which is resource-consuming. In addition, we found out that healthcare datasets are relatively small compared to other datasets, and can have errors, inconsistencies, and other quality issues that greatly affect the accuracy of machine learning models, making the test results unreliable.

Additionally, we were only able to simulate a maximum of four clients due to GPU limitations. In an ideal scenario, we would have been able to work with thousands of clients, as this would have provided us with a more representative sample of data. The limited number of clients also impacted the scalability of our proposed technology.

Furthermore, we faced limitations in data storage and amount. The use of homomorphic techniques greatly increased the size of the images, which resulted in significant storage requirements. Additionally, we only had access to 45 GB of storage, which further limited the amount of data we could work with.

6.2 Future Works

Our future work will be focused on enriching the categories of the data. For now, we have only worked on pituitary tumor classification. In the future, we want to bring heart disease, pneumonia detection, fracture detection, and all of the medical tests that can be done using image data.

Another focus of ours will be on heterogeneous models. So far, the users need to choose the model. We will work on model fusion which will allow us to fusion all of the models and get the best from every model.

We have used partially homomorphic encryption in our advanced data fabric architecture. This encryption model restricts our model efficiency. In the future, we want to take the necessary preprocessing steps which will ensure a better outcome.

Bibliography

[1] *Public law 104 - 191 - health insurance portability and accountability act of 1996*, en, https://www.govinfo.gov/app/details/PLAW-104publ191.

[2] *Council of european union, 2018 reform of eu data protection rules, european commission.*

[3] L. Ponemon, "Third annual benchmark study on patient privacy & data security," *Ponemon Institute, Traverse City, MI*, 2012.

[4] H. Kupwade Patil and R. Seshadri, "Big data security and privacy issues in healthcare," in *2014 IEEE International Congress on Big Data*, Anchorage, AK: IEEE, Jun. 2014.

[5] *Definition of data fabric - gartner information technology glossary.* [Online]. Available: https://www.gartner.com/en/information-technology/glossary/data-fabric.

[6] K. Bonawitz, H. Eichner, W. Grieskamp, *et al.*, "Towards federated learning at scale: System design," *Proceedings of machine learning and systems*, vol. 1, pp. 374–388, 2019.

[7] L. Morris, "Analysis of partially and fully homomorphic encryption," *Rochester Institute of Technology*, pp. 1–5, 2013.

[8] N. G. Kuftinova, O. I. Maksimychev, A. V. Ostroukh, A. V. Volosova, and E. N. Matukhina, "Data fabric as an effective method of data management in traffic and road systems," in *2022 Systems of Signals Generating and Processing in the Field of on Board Communications*, Moscow, Russian Federation: IEEE, Mar. 2022.

[9] V. Theodorou, I. Gerostathopoulos, I. Alshabani, A. Abelló, and D. Breitgand, "MEDAL: An AI-driven data fabric concept for elastic cloud-to-edge intelligence," in *Advanced Information Networking and Applications*, ser. Lecture notes in networks and systems, Cham: Springer International Publishing, 2021, pp. 561–571.

[10] C. Stamatellis, P. Papadopoulos, N. Pitropakis, S. Katsikas, and W. J. Buchanan, "A privacy-preserving healthcare framework using hyperledger fabric," *Sensors*, vol. 20, no. 22, 2020, ISSN: 1424-8220. DOI: 10.3390/s20226587. [Online]. Available: https://www.mdpi.com/1424-8220/20/22/6587.

[11] A. Roehrs, C. A. da Costa, and R. da Rosa Righi, "Omniphr: A distributed architecture model to integrate personal health records," *Journal of Biomedical Informatics*, vol. 71, pp. 70–81, 2017, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2017.05.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046417301089.

[12] Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, and M. Guizani, "Medshare: Trust-less medical data sharing among cloud service providers via blockchain," *IEEE Access*, vol. 5, pp. 14 757–14 767, 2017. DOI: 10.1109/ACCESS.2017.2730843.

[13] Y. Ming and T. Zhang, "Efficient privacy-preserving access control scheme in electronic health records system," *Sensors*, vol. 18, no. 10, p. 3520, 2018.

[14] J. Fu, J. Xu, S. Zhang, and C. Zhang, "Research and design of square kilometer array astronomical data management model based on fabric," in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, Rhodes, Greece: IEEE, Nov. 2020.

[15] D. Shin and S. A. Kim, "Designing through simulation: Integration of ubiquitous street facilities into the urban media fabric," in *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, Gyeongju, South Korea: IEEE, Sep. 2008.

[16] P. Kalyani, M. Masooda, and P. Namrata, "Preserving privacy of data in distributed systems using homomorphic encryption," in *Advances in Intelligent Systems and Computing*, ser. Advances in intelligent systems and computing, Singapore: Springer Singapore, 2021, pp. 306–313.

[17] K. Munjal and R. Bhatia, "A systematic review of homomorphic encryption and its contributions in healthcare industry," en, *Complex intell. syst.*, pp. 1–28, May 2022.

[18] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning," in *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020)*, 2020.

[19] F. Wibawa, F. O. Catak, M. Kuzlu, S. Sarp, and U. Cali, "Homomorphic encryption and federated learning based privacy-preserving cnn training: Covid-19 detection use-case," in *Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference*, 2022, pp. 85–90.

[20] J. Park and H. Lim, "Privacy-preserving federated learning using homomorphic encryption," *Applied Sciences*, vol. 12, no. 2, p. 734, 2022.

[21] A. Madi, O. Stan, A. Mayoue, A. Grivet-Sébert, C. Gouy-Pailler, and R. Sirdey, "A secure federated learning framework using homomorphic encryption and verifiable computing," in *2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS)*, IEEE, 2021, pp. 1–8.

[22] W. Ou, J. Zeng, Z. Guo, W. Yan, D. Liu, and S. Fuentes, "A homomorphic-encryption-based vertical federated learning scheme for risk management," *Computer Science and Information Systems*, vol. 17, no. 3, pp. 819–834, 2020.

[23] I. Kotsiuba, A. Velvkzhanin, Y. Yanovich, I. S. Bandurova, Y. Dyachenko, and V. Zhygulin, "Decentralized e-health architecture for boosting healthcare analytics," in *2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, IEEE, 2018, pp. 113–118.

[24] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: Preserving security and privacy," en, *J. Big Data*, vol. 5, no. 1, Dec. 2018.

[25] C. Zhang, R. Ma, S. Sun, Y. Li, Y. Wang, and Z. Yan, "Optimizing the electronic health records through big data analytics: A knowledge-based view," *IEEE Access*, vol. 7, pp. 136 223–136 231, 2019.

[26] N. Janbi, R. Mehmood, I. Katib, A. Albeshri, J. M. Corchado, and T. Yigitcanlar, "Imtidad: A reference architecture and a case study on developing distributed AI services for skin disease diagnosis over cloud, fog and edge," en, *Sensors (Basel)*, vol. 22, no. 5, p. 1854, Feb. 2022.

[27] J. Morley, C. C. V. Machado, C. Burr, *et al.*, "The ethics of AI in health care: A mapping review," en, *Soc. Sci. Med.*, vol. 260, no. 113172, p. 113 172, Sep. 2020.

[28] I. Gupta, *Decentralization of artificial intelligence: Analyzing developments in decentralized learning and distributed AI networks*, 2020.

[29] G. A. Montes and B. Goertzel, "Distributed, decentralized, and democratized artificial intelligence," en, *Technol. Forecast. Soc. Change*, vol. 141, pp. 354–358, Dec. 2018.

[30] Q.-S. Hua, H. Fan, M. Ai, *et al.*, "Nearly optimal distributed algorithm for computing betweenness centrality," in *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, Nara, Japan: IEEE, Jun. 2016.

[31] L. C. Coronado-Garcia, J. A. Gonzalez-Fuentes, P. J. Hernandez-Torres, and C. Perez-Leguizamo, "An autonomous decentralized system architecture using a software-based secure data field," in *2011 Tenth International Symposium on Autonomous Decentralized Systems*, Tokyo, Japan: IEEE, Mar. 2011.

[32] F. Rahman, M. Slepian, and A. Mitra, "A novel big-data processing framwork for healthcare applications: Big-data-healthcare-in-a-box," in *2016 IEEE International Conference on Big Data (Big Data)*, Washington DC,USA: IEEE, Dec. 2016.

[33] A. M.-H. Kuo, "Opportunities and challenges of cloud computing to improve health care services," en, *J. Med. Internet Res.*, vol. 13, no. 3, e67, Sep. 2011.

[34] M. J. Kaur and V. P. Mishra, "Analysis of big data cloud computing environment on healthcare organizations by implementing hadoop clusters," in *2018 Fifth HCT Information Technology Trends (ITT)*, Dubai, United Arab Emirates: IEEE, Nov. 2018.

[35] V. K. Nigam and S. Bhatia, "Impact of cloud computing on health care," *Int Res J Eng Technol*, vol. 3, no. 5, pp. 2804–2810, 2016.

[36] Y. Guo, M.-H. Kuo, and T. Sahama, "Cloud computing for healthcare research information sharing," in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, Taipei, Taiwan: IEEE, Dec. 2012.

[37] F. Gao, S. Thiebes, and A. Sunyaev, "Rethinking the meaning of cloud computing for health care: A taxonomic perspective and future research directions," en, vol. 20, no. 7, e10041, Jul. 2018.

[38] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren, and A. Alamri, "A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography," *IEEE Access*, vol. 5, pp. 22 313–22 328, 2017.

[39] T. D. Dang, D. Hoang, and D. N. Nguyen, "Trust-based scheduling framework for big data processing with MapReduce," *IEEE trans. serv. comput.*, vol. 15, no. 1, pp. 279–293, Jan. 2022.

[40] M. Nickparvar, *Brain tumor mri dataset*, 2021. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

[41] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.